УДК 004.93

С.А. Субботин
Запорожский национальный технический университет
Украина, 69063, г. Запорожье, ул. Жуковского, 64

Формирование выборок с пространственной локализацией и преобразованием на обобщённую ось

S.A. Subbotin
Zaporizhzhya National Technical University
Ukraine, 69063, Zaporizhzhya, Zhukovskiy st., 64

Sample Formation with Spatial Localization and Transformation to the Generalized Axis

С.А. Субботін
Запорізький національний технічний університет
Україна, 69063, м. Запорожжя, вул. Жуковського, 64

Формування вибірок із просторовою локалізацією та перетворенням на узагальнену вісь

В статье предложено новое решение актуальной научно-практической задачи формирования выборок для автоматизации классификации данных. Впервые предложен метод формирования выборок, который осуществляет иерархическую обработку выборки данных порционно и проецирует данные на обобщённую ось с учётом их глобальной и локальной топологии, что позволяет существенно сократить объём выборки и существенно уменьшает требования к ресурсам ЭВМ.
Ключевые слова: выборка, диагностирование, распознавание, анализ данных.

The new decision of an actual scientific and practical task of sample formation to automate data classification is proposed in the paper. The method of sample forming with a hierarchical data sample processing by portions and projecting the data on the generalized axis according to their global and local topology is firstly proposed. It allows to significantly reduce the size of the sample and significantly reduces the resource requirements of a computer.
Key words: sample, diagnosis, pattern recognition, data analysis.

У статті запропоновано новий розв’язок актуальної науково-практичної задачі формування вибірок для автоматизації класифікації даних. Вперше запропоновано метод формування вибірок, який здійснює ієрархічну обробку вибірки даних порційно і проектує дані на узагальнену вісь з урахуванням їх глобальної та локальної топології, що дозволяє істотно скоротити обсяг вибірки та істотно зменшує вимоги до ресурсів ЕОМ.
Ключові слова: вибірка, діагностування, розпізнавання, аналіз даних.

Введение

Синтез диагностических и распознающих моделей на основе методов вычислительного интеллекта в ряде прикладных задач предполагает необходимость оперировать выборками данных большого объема. Это влечёт за собой существенные затраты

времени на обработку данных, а также требует наличия значительных объемов оперативной и дисковой памяти ЭВМ. Поэтому актуальной задачей является сокращение размерности выборок данных [1-5].

Традиционным и наиболее широко применяемым подходом при решении данной задачи является использование методов отбора информативных признаков [1-5], которые удаляют из исходного набора наименее информативные признаки, и методов конструирования признаков [5], [6], которые заменяют исходный набор признаков рассчитанным на его основе набором искусственных признаков меньшего размера.

Однако, если изначально заданный набор признаков не является избыточным либо объем выборки (число экземпляров в ней) чрезвычайно велик для представления и обработки в памяти ЭВМ, применение этих методов оказывается чрезвычайно затруднительным, а результаты их работы либо приводят к потере существенной для дальнейшего анализа информации, либо не позволяют сохранить исходную интерпретабельность данных.

Другим, существенно реже используемым на практике, подходом при решении данной задачи является сокращение объёма выборки. Как правило, это реализуется посредством извлечения случайных подвыборок из исходной выборки [7-9], что может приводить к формированию нерепрезентативных в топологическом смысле выборок вследствие невключения в них редко встречающихся экземпляров на границах классов, представленных в исходной выборке.

В [10-13] автором предложены переборные и эволюционные методы формирования выборок, а также модель (комплекс критериев) качества выборки, которые позволяют обеспечить формирование из исходной выборки подвыборок меньшего объема, обладающих в системе используемых критериев наилучшими свойствами. Однако для выборок очень большого объема применение данных методов и модели оказывается весьма затратным как с вычислительной точки зрения, так и с точки зрения ресурсов оперативной и дисковой памяти.

Целью данной работы является создание метода формирования и редукции выборок, позволяющего обрабатывать исходные выборки большого объема.

Постановка задачи

Пусть мы имеем исходную выборку $X = \langle x, y \rangle$ – набор S прецедентов о зависимости $y(x)$, $x = \{x^s\}$, $y=\{y^s\}$, $s = 1, 2, ..., S$, характеризующихся набором N входных признаков $\{x_j\}$, $j = 1, 2, ..., N$, где j – номер признака, и выходным признаком y.

Каждый s-й прецедент представим как $\langle x^s, y^s \rangle$, $x^s=\{x^s_j\}$, где x^s_j – значение j-го входного, а y^s – значение выходного признака для s-го прецедента (экземпляра) выборки, $y^s \in \{1, 2, ..., K\}$, где K – число классов, $K>1$.

Тогда задача формирования обучающей выборки может быть представлена как задача выделения из исходной выборки $X = \langle x, y \rangle$ подвыборки X^*, $X^* \subset X$, меньшего объёма $S^*<S$, обладающей наиболее важными свойствами исходной выборки.

Поскольку для задач автоматизации классификации данных наиболее важным является сохранение топологии классов, то формируемая подвыборка должна обеспечивать сохранение экземпляров исходной выборки, находящихся на границах классов.

Метод формирования выборок

Для обнаружения экземпляров, находящихся на границах классов, в общем случае необходимо решить задачу кластер-анализа, что требует определения расстояний между всеми экземплярами выборки. Это, в свою очередь, требует либо загрузки

всей выборки в память ЭВМ (что не всегда возможно из-за ограниченного объёма оперативной памяти), либо многократных проходов по исходной выборке (что вызывает значительные затраты машинного времени), а также приводит к необходимости хранить и обрабатывать матрицу расстояний между экземплярами большой размерности.

Для устранения отмеченных недостатков предлагается заменить обработку экземпляров на обработку их описаний в виде числовых скаляров, которые характеризуют положение экземпляров в пространстве признаков. При этом, заменив экземпляры, характеризующиеся N признаками, на представления в виде скаляров, мы отобразим N-мерное пространство признаков в одномерное пространство.

Исходная выборка, будучи отображённой в одномерное пространство, позволит выделить на одномерной оси интервалы её значений, соответствующие кластерам разных классов в исходном N-мерном пространстве. Определив границы интервалов на одномерной оси, можно найти ближайшие к ним экземпляры, которые и составят формируемую подвыборку.

Преобразование экземпляров исходной выборки на обобщённую ось предлагается осуществлять с двухуровневой иерархией: в начале блоков-кластеров относительно координат в пространстве интервалов признаков, а затем внутри блоков относительно координат в пространстве признаков. Такое преобразование позволит лучше сохранить топологию распределения классов в пространстве признаков.

Поскольку в процессе отбора экземпляров для формирования выборок необходимо выполнять весьма трудоёмкие по времени операции сортировки экземпляров, для сокращения затрат времени предлагается выполнять сортировку и отбор экземпляров небольшими группами отдельно для каждого блока-кластера, после чего объединять результаты таких обработок в обучающую выборку.

Приведенные выше идеи лежат в основе предлагаемого метода.

Этап определения характеристик исходной выборки. Просматривая экземпляры исходной выборки X найти минимальные и максимальные значения для каждого j-го признака x_j: $x_j^{\min} = \min\limits_{s=1,2,...,S}\{x_j^s\}$, $x_j^{\max} = \max\limits_{s=1,2,...,S}\{x_j^s\}$, $j = 1, 2, ..., N$.

Этап формирования разбиения пространства признаков. Вначале необходимо определить k – число интервалов для разбиения оси каждого признака и Q – число прямоугольных областей в пространстве N признаков.

Очевидно, что, с одной стороны, число кластеров-областей в пространстве признаков Q не может быть меньше числа классов K.

С другой стороны, число областей Q должно быть меньше числа экземпляров в исходной выборке S.

Число интервалов, на которые разбиваются оси значений признаков, для регулярного разбиения может быть определено как $k = \sqrt[N]{Q}$, но не может быть меньше двух.

При этом k и Q должны быть целыми положительными числами.

Таким образом, исходя из того, что $K \le Q < S$ и $2 \le k^N = Q$, эвристически зададим правило для определения числа интервалов регулярного разбиения пространства признаков:

$$k = \begin{cases} \left\lceil e^{\frac{\ln \alpha S}{N}} \right\rceil, \text{если} \left\lceil e^{\frac{\ln \alpha S}{N}} \right\rceil > \max\left\{\left\lceil \sqrt[N]{K} \right\rceil, 2\right\}, \\ 2, \text{иначе,} \end{cases}$$

где α – заданная константа, регулирующая число формируемых областей, $0<\alpha<1$. Для малых выборок целесообразно задавать α = 0,8 ... 0,9, а для больших – α = 0,3 ... 0,5.

Далее следует сформировать массив q, сопоставляющий сочетанию номеров интервалов признаков $\{k_j\}$, где k_j – номер интервала значений по j-у признаку, номера прямоугольных блоков-кластеров в пространстве признаков $q(\{k_j\})$.

Этап отображения экземпляров выборки на обобщённую ось. Для каждого s-го экземпляра исходной выборки $<x^s, y^s>$, s=1, 2, ..., S:

– по каждому j-у признаку x_j определить номер интервала значений признака, в который попадает текущий экземпляр:

$$k_j^s = \begin{cases} \left\lfloor \beta_j (x_j^s - x_j^{\min}) \right\rfloor, \beta_j (x_j^s - x_j^{\min}) - \left\lfloor \beta_j (x_j^s - x_j^{\min}) \right\rfloor \geq 0{,}5; \\ \left\lceil \beta_j (x_j^s - x_j^{\min}) + 0{,}5 \right\rceil, \text{иначе}, \end{cases}$$

где $\beta_j = \dfrac{k}{x_j^{\max} - x_j^{\min}}$;

– определить номер области (прямоугольного блока) в исходном пространстве признаков, в которую попадает s-й экземпляр: $q^s = q(\{k_j^s\})$;

– определить координату экземпляра по обобщённой оси:

$$x_*^s = \sum_{j=1}^{N} (k_j^s - 1)^2 + \frac{1}{\pi} \arccos\left(\frac{\sum_{j=1}^{N} k_j^s}{\sqrt{N \sum_{j=1}^{N} \left(k_j^s\right)^2}} \right) + \frac{1}{k} \sum_{j=1}^{N} (\beta_j (x_j^{\min} + (k_j^s - 1)\beta_j^{-1} - x_j^s))^2 +$$

$$+ \frac{1}{k\pi} \arccos\left(\frac{\sum_{j=1}^{N} x_j^s (x_j^{\min} + (k_j^s - 1)\beta_j^{-1})}{\sqrt{\sum_{j=1}^{N} \left(x_j^s\right)^2} \sqrt{\sum_{j=1}^{N} \left(x_j^{\min} + (k_j^s - 1)\beta_j^{-1}\right)^2}} \right).$$

Первый компонент данного преобразования определяет квадрат расстояния блока текущего экземпляра от начала координат в пространстве номеров интервалов признаков, второй компонент определяет взвешенный угол между блоком текущего экземпляра и началом координат в пространстве номеров интервалов признаков, третий компонент определяет взвешенный квадрат нормированного расстояния от текущего экземпляра до начала координат внутри блока-кластера, к которому он принадлежит, в пространстве признаков, а четвёртый компонент – взвешенный угол между экземпляром и началом координат в пространстве признаков;

– занести в подвыборку Ω^{q^s} для соответствующей области q^s текущий экземпляр с координатой по обобщённой оси в виде кортежа $<x^s_*, s, y^s>$: $\Omega^{q^s} = \Omega^{q^s} \cup <x^s_*, s, y^s>$.

Этап анализа экземпляров кластеров по обобщённой оси. Последовательно для каждой p-й области Ω^p, p = 1, ,2, ..., Q:

– упорядочить экземпляры-кортежи $<x^s_*, s, y^s>$ по возрастанию координаты на обобщённой оси;

– просматривая экземпляры-кортежи $<x^s_*, s, y^s>$ p-й области на обобщённой оси слева направо (от меньших значений координаты по обобщённой оси к большим), выполнять отбор экземпляров: если текущий экземпляр-кортеж $<x^s_*, s, y^s>$ является крайним слева или справа по обобщённой оси для данного блока, либо если ближайшие к экземпляру левый и правый экземпляры принадлежат к разным классам, то включить его с добавлением номера области экземпляра в набор кортежей обучающей выборки:

$$\Omega_{об.} = \Omega_{об.} \cup \{ <x^s{}_*, s, y^s, q^s> \mid \neg\exists x_*^g : x_*^g \in \Omega^p, g \neq s, (x_*^g < x_*^s) \vee (x_*^g > x_*^s) \},$$

$$\Omega_{об.} = \Omega_{об.} \cup \{ < x_*^s, s, y^s, q^s > \mid \exists x_*^l, x_*^r \in \Omega^p, l \neq r : x_*^l < x_*^s, x_*^r > x_*^s, \tau_l = 1, \tau_r = 1 \},$$

$$\tau_l = \begin{cases} 1, | x_*^l - x_*^s | \leq | x_*^a - x_*^s |, y^s = y^l, \forall x_*^a : x_*^a \in \Omega^p, a \neq s, x_*^a < x_*^s; \\ 0, \text{иначе}, \end{cases}$$

$$\tau_r = \begin{cases} 1, | x_*^r - x_*^s | \leq | x_*^a - x_*^s |, y^s = y^r, \forall x_*^a : x_*^a \in \Omega^p, a \neq s, x_*^a > x_*^s; \\ 0, \text{иначе}. \end{cases}$$

Этап анализа экземпляров обучающей выборки по обобщённой оси. Для экземпляров-кортежей набора кортежей обучающей выборки $\Omega_{об.}$ выполнять:

– упорядочить экземпляры-кортежи $<x^s{}_*, s, y^s, q^s>$ по возрастанию координаты на обобщённой оси $x^s{}_*$;

– просматривая экземпляры-кортежи на обобщённой оси слева направо (от меньших значений координаты по обобщённой оси к большим) выполнять отбор экземпляров: если текущий экземпляр-кортеж не является крайним слева или справа и ближайшие к экземпляру левый и правый экземпляры принадлежат к тому же классу, то исключить экземпляр из обучающей выборки.

Упорядочить экземпляры-кортежи обучающей выборки по возрастанию номера экземпляра в исходной выборке.

Этап формирования обучающей и тестовой выборок.

Просматривая исходную выборку X, поместить текущий экземпляр $<x^s, y^s>$ в обучающую выборку X^*,если его номер содержится в $\Omega_{об.}$: $X^* = X^* \cup \{<x^s, y^s> \mid \exists <x^p{}_*, s, y^p, q^s> \in \Omega_{об.}\}$, в противном случае – поместить экземпляр $<x^s, y^s>$ в тестовую выборку X':$X' = X' \cup \{<x^s, y^s> \mid \neg\exists <x^p{}_*, s, y^p, q^s> \in \Omega_{об.}\}$.

Анализ сложности метода

Для определения целесообразности применения предложенного метода для конкретной задачи на практике, используя нотацию Ландау в так называемом «мягком виде», оценим сложность этапов предложенного метода.

Временная сложность предложенного метода составит $O(3S^2N + NkQ(S + 1) + 29NS + 6Q + 4S + S\ln(S^2Q^{-1}))$, а пространственная сложность – $O(3N + S + NQk + 3SQ)$.

Для упрощения оценок сложности метода введем следующие допущения. Поскольку N<<S, примем, например, $N = 0{,}01S$.

Примем также $K = 2$, $\alpha = 0{,}5$.

Тогда $k = e^{\frac{\ln(0{,}5S)}{N}} = (0{,}5S)^{N^{-1}}$, $Q = k^N$=0,5S. С учётом принятых допущений получим оценки сложности метода: временной – $O(0{,}03S^3+0{,}005S^2(0{,}5S)^{\frac{100}{S}}(S+1)+0{,}29S^2+S(7+\ln(2S)))$, пространственной – $O(1{,}03S+0{,}005S^2(0{,}5S)^{\frac{100}{S}}+1{,}5S^2)$.

Поскольку S>>100, с запасом примем $(0{,}5S)^{\frac{100}{S}} \approx 2$, тогда получим оценки сложности метода: временной – $O(0{,}04S^3+0{,}3S^2+7S+S\ln(2S))$, пространственной – $O(1{,}03S+1{,}51S^2)$.

Обозначим размерность обучающей выборки $n = NS \approx 0{,}01S^2$.

Тогда, с учётом принятых допущений, округляя, получим оценки сложности метода: временной – $O(40n\sqrt{n}+30n+100\sqrt{n}+0,5\ln n)$, пространственной – $O(151n+10,3\sqrt{n})$.

Для предложенного метода представляет практический интерес определить, насколько обработка исходной выборки с разбиением на области-кластеры позволяет сократить затраты времени на сортировку экземпляров по сравнению с сортировкой всех экземпляров выборки по обобщённой оси.

При использовании эффективного метода сортировки вычислительную сложность сортировки всей выборки по обобщённой оси можно оценить как $O(S\ln S)$, вычислительную сложность обработки экземпляров подвыборки можно оценить как $O\left(\frac{S}{Q}\ln\frac{S}{Q}\right)$, тогда вычислительную сложность сортировки всех подвыборок для кластеров можно оценить как $O\left(S\ln\frac{S}{Q}\right)$.

В результате можно оценить, во сколько раз повышается скорость сортировки экземпляров с разбиением на области-кластеры по сравнению с сортировкой всех экземпляров выборки по обобщённой оси и определить критерий выбора числа кластеров:

$$\begin{cases} F = \dfrac{S\ln S}{S\ln\dfrac{S}{Q}} = \log_{\frac{S}{Q}} S \to \max, \\ Q \geq K, \sqrt[N]{Q} \geq 2. \end{cases}$$

Поскольку на практике значение S является фиксированным (заданным), то задача оптимизации времени сортировки сводится к определению такого значения Q, при котором достигается максимум F при заданных ограничениях.

Эксперименты и результаты

Для экспериментальной проверки работоспособности предложенного метода была разработана его программная реализация, с помощью которой проводились эксперименты по сокращению объема реальных выборок данных для различных практических задач [14-16], а также синтетических выборок, сгенерированных по заданным правилам, характеристики которых приведены в табл. 1.

Таблица 1 – Характеристики задач и сформированных выборок

Задача / выборка	K	N	S^*/S
Классификация автотранспортных средств по изображению [14]	2	26	0,13
Диагностирование патологий плода по кардиотокограмме [15]	3	23	0,09
Предсказание типа лесного покрова [16]	7	54	0,08
Синтетическая выборка 1	2	10	0,14
Синтетическая выборка 2	2	50	0,12
Синтетическая выборка 3	2	100	0,13

Результаты проведенных экспериментов подтвердили работоспособность и практическую применимость предложенного метода, а также программного обеспечения, реализующего его.

Как видно из табл. 1, использование предложенного метода позволяет в среднем в 8 – 9 раз сократить объём выборки, не требуя при этом загрузки в память ЭВМ исходной выборки, а также многочисленных проходов по исходной выборке, что сущест-

венно снижает требования к ресурсам ЭВМ, обеспечивая при этом сохранение в сформированной подвыборке важнейших для последующего анализа топологических свойств исходной выборки.

Проведенные эксперименты также показали, что использование в предложенном методе разбиения выборки на подвыборки для кластеров позволяет получать выигрыш в скорости сортировки по обобщённой оси в 14 – 20 раз по сравнению с сортировкой всей выборки.

Выводы

В статье предложено новое решение актуальной научно-практической задачи формирования выборок для автоматизации классификации данных.

Научная новизна результатов работы заключается в том, что впервые предложен метод формирования выборок, который осуществляет иерархическую обработку выборки данных порционно и проецирует данные на обобщённую ось с учётом их глобальной и локальной топологии, не требуя при этом загрузки в память ЭВМ исходной выборки, а также многочисленных проходов по исходной выборке, что позволяет существенно сократить объём выборки, существенно уменьшает требования к ресурсам ЭВМ.

Практическая значимость результатов работы состоит в том, что разработано программное обеспечение, реализующее предложенный метод формирования выборок, а также проведены эксперименты по их исследованию при решении практических задач, результаты которых позволяют рекомендовать разработанный метод для использования на практике при решении задач интеллектуального анализа данных.

Дальнейшие исследования могут быть сосредоточены на разработке новых способов формирования описаний экземпляров в виде обобщённых показателей, разработке реализаций предложенного метода для параллельных вычислительных систем и распределенной обработки данных.

Работа выполнена в рамках госбюджетной научно-исследовательской темы Запорожского национального технического университета «Интеллектуальные информационные технологии автоматизации проектирования, моделирования, управления и диагностирования производственных процессов и систем».

Литература

1. Олійник А.О. Інтелектуальний аналіз даних : [навчальний посібник] / Олійник А.О., Субботін С.О., Олійник О.О. – Запоріжжя : ЗНТУ, 2012. – 271 с.
2. Рутковская Д. Нейронные сети, генетические алгоритмы и нечёткие системы / Рутковская Д., Пилинский М., Рутковский Л. ; [пер. с польск. И.Д. Рудинского]. – М. : Горячая линия – Телеком, 2004. – 452 с.
3. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов : [монография] / [Субботин С.А., Олейник Ан.А., Гофман Е.А. и др.] ; под ред. С.А. Субботина. – Харьков : ООО «Компания Смит», 2012. – 317 с.
4. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей : [монография] / [Богуслаев А.В., Олейник Ал.А., Олейник Ан.А. и др.] ; под ред. Д.В. Павленко, С.А. Субботина. – Запорожье : ОАО «Мотор Сич», 2009. – 468 с.
5. Субботин С.А. Формирование выборок и анализ качества моделей на основе нейронных и нейронечётких сетей в задачах диагностики и распознавания образов: [монография]. / Субботин С.А. – Saarbrücken : LAP Lambert academic publishing, 2012. – 232 с. – (ISBN 978-3-8473-4471-1).
6. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken : John Wiley & Sons, 2008. – 339 p.
7. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p.

8. Encyclopedia of survey research methods / [ed. P.J. Lavrakas]. – Thousand Oaks : Sage Publications, 2008. – Vol. 1 – 2. – 968 p.
9. Кокрен У. Методы выборочного исследования / Кокрен У.; [пер. с англ. И.М. Сонина ; под ред. А.Г. Волкова, Н.К. Дружинина]. – М. : Статистика, 1976. – 440 с.
10. Subbotin S.A. The training set quality measures for neural network learning / S.A. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2010. – Vol. 19, № 2. – P. 126-139.
11. Субботин С.А. Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов / С.А. Субботин // Математичні машини і системи. – 2010. – № 1. – С. 25-39.
12. Субботин С.А. Критерии индивидуальной информативности и методы отбора экземпляров для построения диагностических и распознающих моделей / С.А. Субботин // Біоніка інтелекту. – 2010. – № 1. – С. 38-42.
13. Субботин С.А. Методы формирования выборок для построения диагностических моделей по прецедентам / С.А. Субботин // Вісник Національного технічного університету «Харківський політехнічний інститут» : зб. наук. праць. – Харків : НТУ «ХПІ», 2011. – № 17. – С. 149-156.
14. Субботин С.А. Синтез нейро-нечётких моделей для выделения и распознавания объектов на сложном фоне по двумерному изображению / С.А. Субботин // Комп'ютерне моделювання та інтелектуальні системи : зб. наук. праць / [за ред. Д.М. Пізи, С.О. Субботіна]. – Запоріжжя : ЗНТУ, 2007. – С. 68-91.
15. Cardiotocography Data Set [Electronic resource]. – Access mode : http://archive.ics.uci.edu/ml/datasets/Cardiotocography.
16. Covertype Data Set [Electronic resource]. – Access mode : http://archive.ics.uci.edu/ml/datasets/Covertype.

Literatura

1. Olijnyk A. O. Intelektual'nyj analiz danyh / [A.O. Olijnyk, S.O. Subbotin, O.O. Olijnyk : navchal'nyj posibnik]. – Zaporizhzhja : ZNTU, 2012. – 271 s.
2. Rutkovskaja D. Nejronnye seti, geneticheskie algoritmy i nechjotkie sistemy / [D. Rutkovskaja, M. Pilin'skij, L. Rutkovskij ; per. s pol'sk. I. D. Rudinskogo]. – M. : Gorjachaja linija – Telekom, 2004. – 452 s.
3. Intellektual'nye informacionnye tehnologii proektirovanija avtomatizirovannyh sistem diagnostirovanija i raspoznavanija obrazov : monografija / [S.A. Subbotin, An.A. Olejnik, E.A. Gofman, S.A. Zajcev, Al.A.Olejnik] ; pod red. S.A. Subbotina. – Har'kov : OOO «Kompanija Smit», 2012. – 317 s.
4. Progressivnye tehnologii modelirovanija, optimizacii i intellektual'noj avtomatizacii etapov zhiznennogo cikla aviacionnyh dvigatelej : monografija / [A.V. Boguslaev, Al.A. Olejnik, An.A. Olejnik et all.] ; pod red. D.V. Pavlenko, S.A. Subbotina. – Zaporozh'e : OAO «Motor Sich», 2009. – 468 s.
5. Subbotin S.A. Formirovanie vyborok i analiz kachestva modelej na osnove nejronnyh i nejro-nechjotkih setej v zadachah diagnostiki i raspoznavanija obrazov : [monografija] / Subbotin S.A. – Saarbrücken : LAP Lambert academic publishing, 2012. – 232 s. – (ISBN 978-3-8473-4471-1).
6. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken : John Wiley & Sons, 2008. – 339 p.
7. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p.
8. Encyclopedia of survey research methods / [ed. P.J. Lavrakas]. – Thousand Oaks : Sage Publications, 2008. – Vol. 1-2. – 968 p.
9. Kokren U. Metody vyborochnogo issledovanija / U. Kokren ; [per. s angl. I.M. Sonina] ; pod red. A.G. Volkova, N.K. Druzhinina. – M. : Statistika, 1976. – 440 s.
10. Subbotin S.A. The training set quality measures for neural network learning / S.A. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2010. – Vol. 19. – № 2. – P. 126-139.
11. Subbotin S.A. Kompleks harakteristik i kriteriev sravnenija obuchajushhih vyborok dlja reshenija zadach diagnostiki i raspoznavanija obrazov / S.A. Subbotin // Matematichni mashini i sistemy. – 2010. – № 1. – S. 25-39.
12. Subbotin S.A. Kriterii individual'noj informativnosti i metody otbora jekzempljarov dlja postroenija diagnosticheskih i raspoznajushhih modelej / S.A. Subbotin // Bionika intelektu. – 2010. – № 1. – S. 38-42.
13. Subbotin S.A. Metody formirovanija vyborok dlja postroenija diagnosticheskih modelej po precedentam / S.A. Subbotin // Visnik Nacional'nogo tehnichnogo universitetu «Harkivs'kij politehnichnyj institut» : zb. nauk. prac'. – Harkiv : NTU «HPI», 2011. – № 17. – C. 149-156.
14. Subbotin S.A. Sintez nejro-nechetkih modelej dlja vydelenija i raspoznavanija ob#ektov na slozhnom fone po dvumernomu izobrazheniju / S.A. Subbotin // Komp'juterne modeljuvannja ta intelektual'ni sistemy : zb. nauk. prac' / [za red. D.M. Pizy, S.O. Subbotina]. – Zaporizhzhja : ZNTU, 2007. – S. 68-91.

15. Cardiotocography Data Set [Electronic resource]. – Access mode : http://archive.ics.uci.edu/ml/datasets/Cardiotocography.
16. Covertype Data Set [Electronic resource]. – Access mode : http://archive.ics.uci.edu/ml/datasets/Covertype.

S.A. Subbotin

Sample Formation with Spatial Localization and Transformation to the Generalized Axis

The synthesis of diagnostic and pattern recognizing models based on the methods of computational intelligence in some applications requires to operate with a large data samples. This entails a significant over-expenditure of time for the processing of data, and require a large amount of memory and disk space of a computer.

The purpose of the paper is to develop a method for the formation and reduction of samples, allowing to handle a large amount of the original sample.

The new solution of actual scientific and practical task of sample formation to automate data classification has been proposed.

The scientific novelty of the work lies in the fact that the method of sample forming with a hierarchical data sample processing by portions and projecting the data on the generalized axis according to their global and local topology and do not require downloading to the computer memory of the original sample, and numerous passages on the original sample is firstly proposed. Method instead of the original examples process their descriptions in the form of numeric scalars that characterize the status of exapmles in the feature space. In this case, an N-dimensional feature space is transformed to one-dimensional space. In the one-dimensional space it can be identified the intervals of the generalized axis corresponding to clusters of different classes in the original N-dimensional space. Examples nearest to the borders of the intervals can be included in the formed sub-sample. It allows to significantly reduce the size of the sample and significantly reduces the resource requirements of a computer.

The practical significance of the work lies in the fact that it has been developed software that implements the proposed method of sample formation and the experiments on their research at practical problem solving has been conducted, results of which allow to recommend the developed method for use in practice at solving data mining problems.

Further research could focus on the development of new methods to form example description in the generalized measures, and on the development of implementation of the method for parallel and distributed computing.

Статья поступила в редакцию 14.02.2013.